IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 June 2024, accepted 24 July 2024, date of publication 29 July 2024, date of current version 8 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435347

RESEARCH ARTICLE

Adversarial Robustness of Vision Transformers Versus Convolutional Neural Networks

KAZIM ALI[1], MUHAMMAD SHAHID BHATTI[2], ATIF SAEED[2], ATIFA ATHAR[2], MOHAMMED A. AL GHAMDI[3], SULTAN H. ALMOTIRI[4], AND SAMINA AKRAM[5]

[1]Punjab Education Department, Government of Punjab, Layyah 54000, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Lahore 54000, Pakistan
[3]Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah 24211, Saudi Arabia
[4]Department of Cybersecurity, College of Computing, Umm Al-Qura University, Makkah 24211, Saudi Arabia
[5]Department of Computer Science, Faculty of Information Technology, University of Central Punjab, Lahore 54000, Pakistan

Corresponding author: Kazim Ali (kazimravian2003@gmail.com)

This work was supported by the Deputyship for Research and Innovation, Ministry of Education, Saudi Arabia, under Project IFP22UQU4250002DSR216.

ABSTRACT Vision Transformers (ViTs) have proved to be a more powerful substitute for Convolutional Neural Networks (CNNs) in various computer vision tasks, using the self-attention approach to gain remarkable results and observations. However, the adversarial robustness of ViTs against adversarial attack methods raises critical questions, and the issues of using these models in security-related applications remain under discussion. This paper presents a novel and systematic approach to evaluate and compare the adversarial robustness of ViTs with CNNs, explicitly concentrating on the image classification problem. We have performed extensive experiments using state-of-the-art adversarial example attacks, such as the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and DeepFool Attack (DFA). The findings of this research study represent that CNNs are more robust against more straightforward attacks such as FGSM. Still, ViTs show excellent resistance against more dangerous attacks like PGD and DFA attack methods. This work provides useful outcomes revealing the advantages and limitations of CNNs and ViTs, which are helpful for further study and applications regarding safer and more effective use of deep learning models of CNNs and ViTs.

INDEX TERMS Vision transformers, convolutional neural networks, clean accuracy, adversarially robustness, adversarial attacks.

I. INTRODUCTION

Transformers are a kind of deep learning model based on self-attention mechanisms which is mainly used for various tasks in the branch of natural language processing [1], and the reliance on many large language processing models transformed by pre-training have obtained outstanding performances on various tasks of NLP [2], [3], [4], [5]. Transformers are now also employed for computer vision undertakings that include image classification, object recognition, and segmentation. Such models are known as vision transformers (ViTs) and work in the same field as convolutional neural networks (CNNs) [6]. After that, ViTs have been applied to solve many tasks in CV and have outperformed the models based on CNN and Recurrent Neural Networks (RNN) [7], [8], [9]. Even though ViT and its variants demonstrate high results in Computer Vision tasks, it is important that they can do so while in the presence of adversarial attacks.

Adversarial robustness is an essential issue in machine learning especially in the case of deep learning models in computer vision [9]. These are inputs to machine learning models which are designed to be slightly manipulated to make the model fail. Such perturbations are usually very small and cannot be detected by a human but could still lead to a bad prediction in the model [10], [11], [12]. In this respect, adversarial robustness is the capability of the considered architectures, namely ViT and CNN, to provide

The associate editor coordinating the review of this manuscript and approving it for publication was Pasquale De Meo.

accurate predictions and remain efficient within the presence of adversarial perturbations. We can observe that ViTs and CNNs are currently applied to different sorts of tasks, ranging from image or speech recognition to self-driving cars; therefore, it is crucial to improve the model robustness that does not jeopardize the accuracy and is resistant to methods of adversarial attacks. It has been found that even the latest ViT and CNN models are easily susceptible to adversarial attacks, which is an issue in security-sensitive contexts [13], [14].

There are two types of adversarial attacks such as White-Box-Attacks, which occur when the attacker has complete knowledge of the model, including its architecture and parameters [15], and Black-Box-Attack: which are performed without any information on the internal architecture of a model and the malicious actors may only see a model's inputs and outputs [16]. The motivation behind this research is (i) Security Implications: As shown by the last examples, such vulnerability presents considerable security threats in numerous fields, ranging from cybersecurity and finance to autonomous systems, and (ii) Understanding Model Behavior: Identifying the reason for such weaknesses can help elicit the models' decision-making processes and the features that make them susceptible to adversarial attacks.

Researchers have also proposed defence techniques to increase the adversarial robustness against adversary method attacks such as adversarial training [17] this involves training a model on a mixture of normal and adversarial examples to improve its resilience. Other defensive techniques such as input preprocessing [18], model regularization [19], and network architecture modifications [20], are explored to defend against adversarial attacks. There are challenges and open questions are still facing adversarial robustness described as (i) **Trade-off Between Accuracy and Robustness**: Increasing a model's robustness often comes at the cost of reduced accuracy on unperturbed data [21], (ii) **Scalability and Efficiency**: Making models robust against adversarial attacks can require additional computational resources and more complex training processes [22], and (iii) **Evolving Nature of Attacks**: As defensive techniques improve, new and more sophisticated adversarial attack methods continue to emerge [23]. There is an ongoing race among researchers to develop more sophisticated adversarial attacks and create robust defence mechanisms. The ultimate goal is to build deep learning systems based on ViTs and CNNs that are not only high-performing but also secure and reliable in real-world scenarios.

In this research study, we analyze the adversarial robustness of ViT and CNN models on the image classification task against the adversarial perturbation methods We have compared the adversarial robustness of ViTs with CNN models. We have clearly described which model is more robust against adversarial perturbation attacks in discussion section VI. The key contribution of this work is as under:

- The present work gives a comprehensive analysis and comparison of the adversarial robustness of Vision Transformers and Convolutional Neural Networks. To this end, the analysis is carried out considering the image classification task, comparing both types of models when subjected to different adversarial attacks.
- The experiments involve all the recommended tests utilizing extreme adversarial attack techniques, including FGSM, PGD, and DFA. These latter are carried out to evaluate the robustness of such models against such adversarial perturbations.
- Thus, it is discovered that CNNs show higher resistance to simpler forms of adversarial attacks such as FGSM while ViTs are more immune to complex attacks like PGD and DFA. Such distinction is highly beneficial in understanding the effectiveness of each model architecture in adversarial conditions.
- As such, this work helps to make the further use of ViTs and CNNs in highly sensitive security applications safer and more effective by exposing these models' adversarial robustness characteristics. It underlines the potential vulnerabilities of such models and the necessity to develop methods for protection against adversarial attacks used in the real-life applications of such models.
- The implementation details of this research are made publicly available on GitHub, ensuring transparency and facilitating further research. The repository can be accessed at https://github.com/kazimravian/Adversarial-Robustness-of-Vision-Transformers-vs-Convolutional-Neural-Networks.

II. RELATED WORKS

The transformer deep learning models have achieved state-of-the-art results on natural language processing (NLP) tasks and discussed their robustness [1]. The researchers have conducted adversarial attack experiments on transformers and pre-trained large language models based on transformers. They have proved that transformer algorithms are more adversarially robust than traditional long short-term memory (LSTM) based on recurrent neural networks (RNN). However, they have only concentrated on the robustness against the adversarial perturbation method in NLP tasks, not on computer vision tasks [24], [25], [26], [27], [28], [29].

During the research, we have found that some work on the adversarial robustness of ViTs and CNNs in computer vision tasks, which we acknowledge their contribution as follows; [30] This paper presents a method to produce adversaries to check the adversarial robustness of the image classification model, [31] has developed a defense strategy which resists against adversarial perturbation methods, [32] presents adversarial training for some text classification problems, [33] Mixup technique train mdels on combined data example to improve adversarial robustness of deep learning models, [34] presents the concept of certified robustness, where a model's robustness is secured within a some certain margin, [35] has give defense technique based gradient masking strategy, [36] discuss the adversarial

training and its variants, [37] proposes the transferability of adversaries among different deep learning based computer vision models, [14] gives a comprehensive study of adversarial examples methods and defences for 3D object detection models, [10] provides a good survey of adversarial attacks and defences in computer vision domain, [38] presents the different adversarial robustness benchmarks and challenges, [39] discusses adversarial attacks and the lack of robustness in pretrained language models like BERT on social media text, [40] has tested the adversarial accuracy against both the white-box and black-box adversarial perturbation methods and analyzed the security of ViTs and CNNs, [41], [42] have presented adversarial robustness of the vision transformers (ViT) against the patch-based adversarial perturbation, [43] develops a pyramid adversarial training method by using data augmentation techniques to improve the robustness of ViTs in the presence of adversarial examples, [44] has studied the transferability of adversarial perturbation among different ViTs, [45], [46], [47], [48], [49] have worked on the robustness of ViTs in different perspective such as to develop defense methods against perturbed data by smoothing techniques, [50] have also discussed the adversarial robustness of other computer vision tasks such as image segmentation and object recognition.

III. TARGET MODELS'S STRUCTURE

We first provide a summary of the structures of target ViTs and CNNs models, which we have used for experiments and results to evaluate the adversarial robustness against adversarial attacks in Tables 1 and 2. These models are trained on the dataset CIFAR10, which has 60,000 images consisting of 10 categories, 50000 for training and 10000 for testing purposes and imagenet/160px-v2 [51], a subset of 10 classified classes from the Imagenet dataset containing 9,469 images for training and 3,925 for testing. Jeremy Howard of FastAI initially prepared it. The objective behind putting together a tiny version of the Imagenet dataset was that running new research ideas/algorithms/experiments on the whole Imagenet takes a lot of time.

Tables 1 and 2 clearly describe the different properties of the structure of models under attack, such as the number of layers, memory size, dense or hidden size, number of trainable parameters, model types (vision transformer or CNN models), and datasets. The clean accuracy of target models on imagenet/160px-v2 and CIFAR10 are shown in Fig. 1 and 2. Clean accuracy means when there is no attack threat.

Fig. 1 shows the accuracy of target models on imagenet/160px-v2. The compared models are: ViT-b16, ViT-b32, ViT-L16, ViT-L32, ResNet50, VGG19, MobileNet, DenseNet.

Fig. 3 shows that DenseNet also has a good accuracy of 0.90, seconded by VGG19 with 0.89. According to Table 7, the least accuracy is recorded by ViT-L32, which is 0.75. Thus, these graphs give a complete and understandable picture of the model performance on the CIFAR10 dataset, revealing the main advantages and disadvantages of target models in clean, accurate conditions.

TABLE 1. Summarize the target ViTs and CNNs model, which are used in experiments.

Model	Layers	Model Size (MB)	Dense Head	Parameters (M)	Type	Dataset
ViT-b16	20	329.70	1000	86.43	Attention-based	imagenette/160px-v2
ViT-b32	20	336.46	1000	88.19	Attention-based	imagenette/160px-v2
ViT-L16	32	1.13	1000	304.14	Attention-based	imagenette/160px-v2
ViT-L32	32	1.14	1000	306.49	Attention-based	imagenette/160px-v2
ResNet50	180	153.86	1024	40.39	CNN-based	imagenette/160px-v2
VGG19	27	84.43	1024	22.13	CNN-based	imagenette/160px-v2
Mobile-Net	91	28.28	1024	7.43	CNN-based	imagenette/160px-v2
Dense-Net	712	99.06	1024	26.20	CNN-based	imagenette/160px-v2

TABLE 2. Describes the structural parameters of target models on the CIFAR10 dataset.

Model	Layers	Model Size (MB)	Dense Head	Parameters (M)	Type	Dataset
ViT-b16	20	329.67	1000	86.42M	Attention-based	CIFAR10
ViT-b32	20	336.41	1000	88.19M	Attention-based	CIFAR10
ViT-L16	20	329.67	1000	86.42M	Attention-based	CIFAR10
ViT-L32	20	336.41	1000	88.19M	Attention-based	CIFAR10
ResNet50	180	97.82	1024	25.07M	CNN-based	CIFAR10
VGG19	27	78.43	1024	20.56	CNN-based	CIFAR10
Mobile-Net	91	16.28	1024	4.29M	CNN-based	CIFAR10
Dense-Net	712	76.56	1024	20.30	CNN-based	CIFAR10

A. VISION TRANSFORMERS (VITS

The transformer model [1] was first introduced in 2017 in the paper "Attention Is All You Need" for sequence-to-sequence problems. It was an alternative to using recurrent or convolutional layers. Vision Transformer (ViT) [6] is a

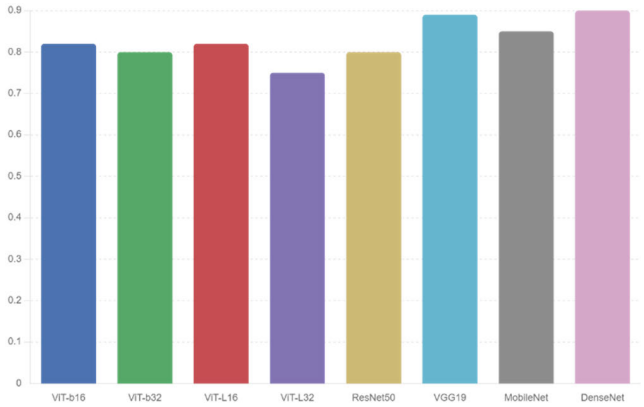


FIGURE 1. Shows the clean accuracy of target models on imagenet/160px-v2 dataset.

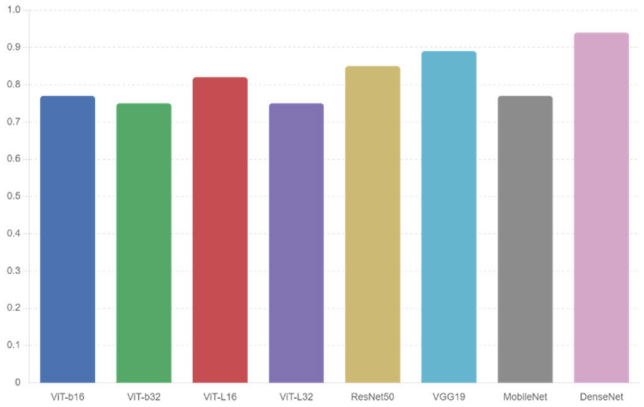


FIGURE 2. Clean accuracy of target model on CIFAR10 dataset.

BERT-style [2] take on transformers for vision, as shown in Figure 2. First, the image is split into smaller patches, and each patch is embedded with a linear projection. The results strongly resemble the token embeddings in BERT, and what follows is virtually identical. The patch embeddings are combined with position embeddings and fed through an ordinary transformer encoder. Some patches are masked or distorted during pretraining, and the objective is to predict the average colour of the mask patch. In the case of 2D images, we can describe them as $x_i \in \mathbb{R}^{\{H\times W\times C\}}(1 \leq i \leq N)$ with resolution $H \times W$ and C channels, it is divided into sequence of $N = \frac{H.W}{P^2}$ flatted 2D patches of size $P \times P, x_i \in \mathbb{R}^{P^2.C}$ $(1 \leq i \leq N)$. The patches are encoded into patch embeddings with a simple convolutional layer, where the kernel size and stride of the convolution is exactly $P \times P$.

B. CNNS MODELS

We have also studied and experimented with the different versions of the CNN models to compare adversarial robustness with vision transformers. These models are ResNet-50, VGG19, Mobile-Net, and Dense-Net, and they are also trained on the imagenet/160px-v2 and CIFAR10 dataset.

IV. ADVERSARIAL ATTACK METHODS

Let us suppose that an image classifier f based on CNNs/ViTs. It takes an image representing the input data and returns an output:

$$f\ (I, \theta) = y$$

The above equation represents the f is an image classifier based on the CNNs/ViTs algorithm to predict the correct category of the original image I, θ represents the trainable parameters of the CNN/ViT model like weights and biases. When the image classifier successfully trains, then the trainable parameters do not change, and the above equation is re-write as follows:

$$f\ (I) = y$$

where I is an image, and y is the prediction of f. The I is an image consisting of raw pixel values belonging to real numbers such that:

$$I_{i,j}\epsilon\mathbb{R}^+$$

For an image classifier, the y is the predicted class, such as "horse" or "truck," derived from a probabilities vector that gets from the output layer of the classifier. Here y is not a vector, but it belongs to a vector of probabilities $1...L$ returned from the model, where L is the number of classes to be predicted. We write as follows:

$$y\epsilon\ \{1 \ldots L\}\ \epsilon\mathbb{R}^+$$

The adversarial examples are developed by the following relation as follows:

$$I^{adv} = I + r$$

where:

- I^{adv}, is the adversarial image example.
- I is the original input data.
- r is the slight change in the original image, also called adversarial perturbations.

For a successful I^{adv}, the result of the image classifier f on I^{adv}, is must different from the result on the original input image I; thus, we can write as follows:

$$f(I^{adv}) \neq f(I)$$

whether the adversarial image examples are generated from adversarial perturbation or patch does not matter. However, the adversarial noise r must be as minimized as possible, which is humanly imperceptible or less perceptible [52]. The adversarial perturbation is calculated by using $L_p - norm$ metric during an adversarial attack and the aim is to find the nearest image of I but the perturbation should be remains as small as possible, which we write as follows:

$$argmin_r\{||r||_p : f(I^{adv}) \neq f(I)\}$$

The attacker tries to get a minimum quantity of r, which satisfies the primary criteria. It can be achieved by using

a constrained optimization algorithm [53], and the need to choose an optimization algorithm is, therefore, to produce an adversarial image example by solving the following relation:

$$I^{adv} = I + argmin_r\{||r||_p : f(I^{adv}) \neq f(I)\}$$

The **Fast Gradient Sign Method (FGSM)** is faster and less computationally capable of creating adversarial image examples [30]. The FGSM increases the loss of the target model, decreases predicted class probabilities, and increases confusion during the target models' classification. The training loss function of the target model will increase, reducing classification confidence and increasing the likelihood of inter-class confusion. FGSM determines the gradient of the loss function w.r.t the input image I, multiplying a constant ε with the sign of gradient to produce adversarial perturbation. The adversarial example is created by solving the following relation:

$$I' = I + \varepsilon.sign(\nabla_I L(I, y))$$

where I and I', are original and adversarial images, respectively. The $\nabla_I L(I, y)$ is the slope of the cost function w.r.t original image example I. The backpropagation algorithm [54] determines the gradient of the cost function.

The **projected Gradient Descent Method (PGD)** [31] creates adversarial examples iteratively using FGSM [55] on the clean input image I_0 by adding a small amount of random perturbation α in the original sample image I. The projection first searches the closest image from hyperplane or decision boundary plane. The PGD finds the adversarial image example which is close to the original image example. The following relation explains it:

$$I^{i+1} = Proj_{I+S}(I^i + \alpha.sign\left(\nabla_{I^i} L\left(\theta, I^i, t\right)\right))$$

where I^{i+1} is the perturbed image after $i+1$ iterations, and S is the negative space or region where the adversarial example lies.

The **Deep Fool Attack (DFA)** was introduced by Moosavi-Dezfooli et al. [56], an untargeted attack method for creating an adversarial example and depends on the l_2 distance measure. The DFA is tried to find the minimum distance between the original input and the descion boundary of the target classifier. Decision boundaries are the boundaries that divide different classes in the hyper-plane created by the classifier. Perturbations push the adversarial sample outside its prediction space to misclassify the example by the target models. The whole DFA process is represented in *Algorithm 1*.

Algorithm 1 The process of creating adversarial examples is done using the DFA method.

Input: input image I, target model f
Output: Perturbation $r^{\wedge}$
Initialize: $I_0 \leftarrow I, i \leftarrow 0$
$whilesign\,(f\,(I_i)) = sign\,(f\,(I_0))\,do$
$r_i \leftarrow -\frac{f(I_i)}{||\nabla f(I_i)||_2^2}\nabla f\,(I_i)$
$I_{i+1} \leftarrow I_i + r_i$
$i + 1 \leftarrow i$
end while
return $r^{\wedge} = \sum_i r_i$

V. PROPOSED METHOD TO EVALUATE ADVERSARIAL ROBUSTNESS

The different steps of the proposed method to evaluate and analyze the adversarial robustness of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) in image classification tasks are shown in Fig. 3. The step-by-step discussion and explanation of each phase of the proposed method are given as follows:

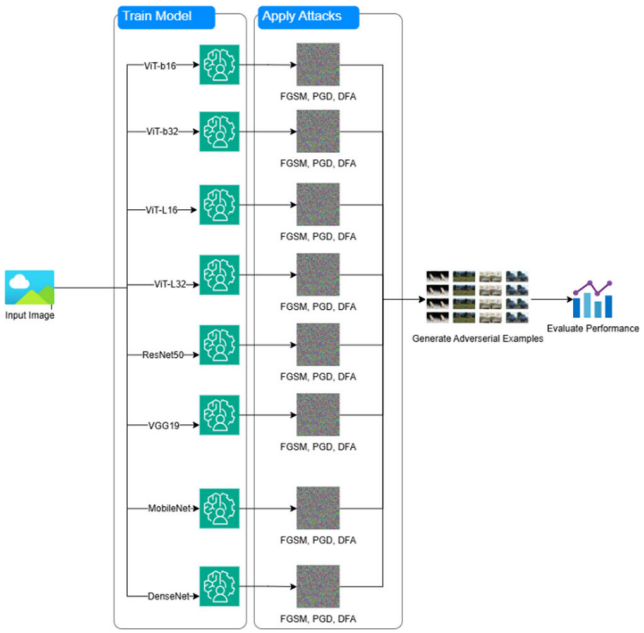


FIGURE 3. Shows the proposed method to validate the adversarial robustness of VITs vs CNNs against adversarial attacks.

A. INPUT IMAGES

The proposed method starts with input images of the data given to the models for evaluating the adversarial robustness against adversarial attacks. These input image data pass through different stages of transformation to test the adversarial robustness of ViTs and CNNs models against adversarial attacks.

B. TARGET VITS AND CNNS MODELS

The model section lists the target models used to validate adversarial robustness. The models are given as follows:

- Vision Transformers (ViTs)
  - Large ViT models include ViT_L32 with 32×32 patch sizes and ViT_L16 with 16×16 patch sizes.
  - Base vision models include ViT_B32 and ViT_B16 with 32×32 and 16×16 patch sizes, respectively.
- Convolutional Neural Networks (CNNs):

- The proposed method has used popular CNN models, which are standard and state-of-the-art for image classification tasks such as DenseNet, MobileNet, ResNet, and VGG19.

C. ADVERSARIAL ATTACKS

This phase of the method consists of adversarial attacks on the input image data. These attacks aim to produce slight perturbations in the images that can fool the models into giving false predictions. The attacks used in this work are as follows:

- **FGSM**: An easy and straightforward attack that creates perturbation by calculating the loss gradient to the input image.
- **PGD**: A repetitive and powerful version of FGSM that creates perturbation in many gradient steps.
- **DFA**: This attack searches for the less perturbation required to alter the prediction of the input sample, iteratively linearizing the model's prediction area.

D. GENERATE ADVERSARIAL EXAMPLES

At this stage, adversarial images are produced to use the attacks mentioned above. These images are perturbed and specially created to fool the models for incorrect predictions.

E. EVALUATE TRAINED MODELS ON ADVERSARIAL EXAMPLES

In this phase, the target model ViTs and CNNs are validated on the adversarial examples created by FGSM, PGD, and DFA attacks. The primary purpose is to observe and analyze how trained models work when receiving the inputs generated to make misclassifications.

F. REPORT ADVERSARIAL ROBUSTNESS

The last phase describes and reports the ups and downs of the adversarial robustness of trained models. It includes analyzing and comparing the performance and efficiency of ViTs and CNNs in the presence of adversarial attacks and also provides which model's architecture is more robust against attacks.

Summarizing the proposed method shows a detailed approach to calculating the adversarial robustness of different trained model architectures of ViT and CNNs. By intentionally using adversarial example attacks and testing the models' efficiency, this research describes the strengths and weaknesses of ViTs and CNNs models in maintaining prediction accuracy under different adversarial situations.

VI. EXPERIMENTS AND RESULTS

This section provides a detailed comparison of the adversarial robustness of target CNN and ViT models against adversarial perturbation attacks such as FGSM, PGD, and DFA. In the end, the empirical results of the final experiments prove which target models are more or less robust against adversarial attacks.

TABLE 3. Shows the various parameters of FGSM, PGD, and DFA attacks like step size, number of, and maximum strength.

Attacks	Step-Size α	Iterations i	Max. Strength ε
FGSM	0.003	not iterative	0.03
PGD	0.001	20	0.01
DFA	1	3	No-strength

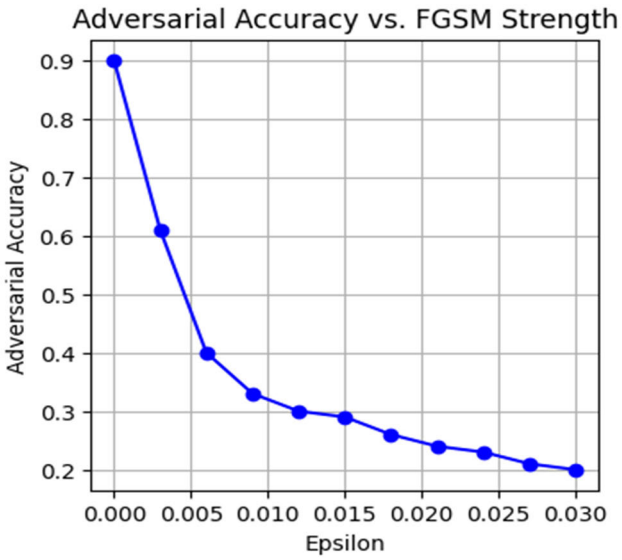


FIGURE 4. Shows the degradation of model accuracy at different perturbation levels during FGSM attack.

A. ADVERSARIAL ATTACKS SETTINGS

The adversarial attack FGSM, PGD, and DFA settings mean setting different parameters such as step size, number of iterations, and maximum attack strength. As mentioned earlier, the values of these parameters are kept the same for all target ViT and CNN models for a fair and clean comparison of adversarial robustness against the attacks. Table 2 shows the different parameters of attacks.

In Table 3, it is concluded that FGSM is a single-step attack. The perturbation is added in the gradient direction with step size = 0.003, not-iterative, and max strength = 0.3, PGD with step-size = 0.001, iterations = 20, and strength = 0.01, DFA with iteration step-size = 3, maximum-iterations = 3. There is no strength value in this type of attack. Here, we also show a few results proving how these adversarial attack parameters gradually degrade the performance (clean accuracy) of the above-mentioned trained models on the mentioned datasets in Fig. 2 and 3. These results were obtained at different levels of perturbation or attack strength and randomly selected during experiments to inspect the effect of adversarial attacks.

Fig. 4 shows the adversarial accuracy at epsilon 0; it is very high, equal to 0.9. However, with increasing epsilon, the accuracy quickly decreases, roughly to 0.2 at epsilon 0.03.

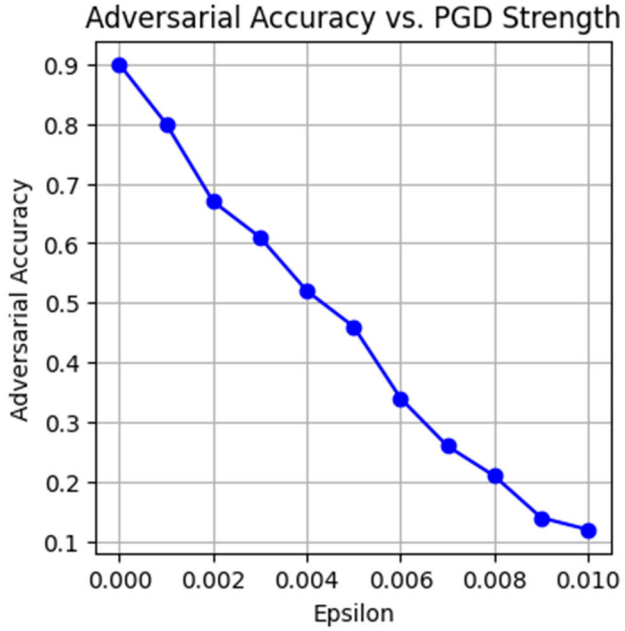


FIGURE 5. Shows the degradation of model accuracy at different perturbation levels during PGD attack.

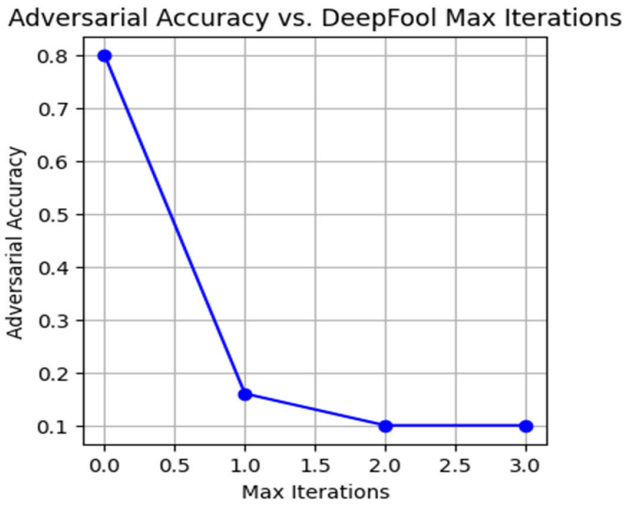


FIGURE 6. Shows the degradation of model accuracy at various perturbation levels during DFA attack.

TABLE 4. Shows clean accuracy, adversarial accuracy, and adversarial robustness (%) of the target (ViTs and CNNs) models against the FGSM perturbation method with maximum strength $\varepsilon = 0.03$ on imagenett dataset.

Target Models	Clean Accuracy	Adversarial Accuracy	Adversarial Robustness(%)
ViT-b16	0.82	0.08	0.09
ViT-b32	0.80	0.11	0.14
ViT-L16	0.82	0.08	0.9
ViT-L32	0.75	0.17	0.22
ResNet50	0.80	0.28	0.35
VGG19	0.89	0.21	0.23
MobileNet	0.85	0.34	0.40
DenseNet	0.90	0.20	0.22

This means that when the FGSM attacks are more potent (higher epsilon values), there is a more significant decrease in the model's accuracy, thus portraying a higher susceptibility to adversarial examples.

Fig. 5 demonstrates the epsilon values ranging from 0.0 to 0.01 with step size 0.001. In the highest epsilon value, 0.01, the accuracy declines from 0.90 to 0.10. It means that as the strength of the attack by PGD increases, the classification capability of the model deteriorates dramatically.

Finally, Fig. 6 shows that at iteration 0, the accuracy is the biggest, approximately equal to 0.80. If the number of iterations increases to 1, the accuracy decreases to about 0.20. As the iterations transit from 1 to 3, the accuracy worsens while levelling off at 0.10. It shows that the accuracy of the adversary considerably reduces once the DeepFool iterations increase and remains low even as more iterations take place, meaning that the DeepFool attack is dangerous in lowering model accuracy even with a few iterations. Some visual examples of adversarial image examples are shown below in Fig. 7.

Fig. 8, Tables 4, 5, and 6, compares the adversarial robustness of various target models—ViT-b16, ViT-b32, ViT-L16, ViT-L32, ResNet50, VGG19, MobileNet, and DenseNet—against three different types of attacks: FGSM, PGD and DFA can be used for penetration testing and to enhance the accuracy of the prediction during the brute force attack. All the models show different degrees of resistance in the various kinds of attack. FGSM Attack: The results of the FGSM attack demonstrate that DenseNet and MobileNet are the most adversarial robust, having scores of approximately 0.40. Again, the result is 0.14, showing good resistance against this style of attack. Similar to VGG19, ResNet50 proved moderately strong with supremacy and deviation values surpassing 0.10. Meanwhile, the measures of ViT models, which are b and L series, are identified as lower as compared to the first one = 0.05 to 0.08. PGD Attack: After being targeted by PGD, all the ViT models, as well as detectors, have improved robustness, especially ViT-L32, ViT-b32, and ViT-b16 with score rates of up to 0.45. VGG19 also appear to be immune, with the proposed architecture having a robustness score of roughly 0.30. At the bottom of the scale, DenseNet and MobileNet demonstrate the weakest results, which are below 0.15; it is shown that DenseNet has the lowest robustness among all the architectures. DFA

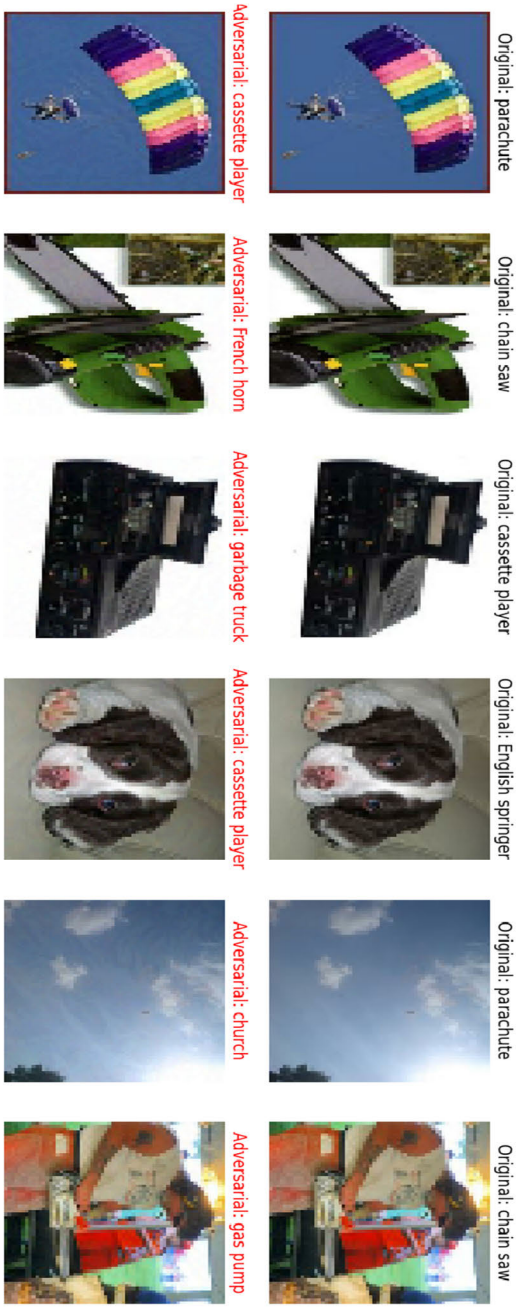


FIGURE 7. Shows original and adversarial images produced by adversarial attacks which are imperceptible by humans.

Attack: In response to DFA attacks, MobileNet and ViT-L32 are the most resistant architectures, with MobileNet at approximately 0.25. Generally, ViT models are pretty stable, specifically ViT-b16 and VT-b32, with a score slightly above 0.20. Again, DenseNet performs the poorest with a score of uttermost robustness, somewhat more than 0.11. It was only 0.5, which is less than the recommended level of 1, thus showing the circuits have a poor resistance to DFA attacks.

To sum up, as can be observed from Table 6, the ViT models are equally or even more robust than other architectures against PGD and DFA attacks. Still, DenseNet and MobileNet are more robust against FGSM attacks. ResNet50 and VGG19 are of moderate robustness against all three types of attacks, which means that the networks are relatively balanced in their adversarial defence.

The comparison of adversarial robustness against FGSM, PGD, and DFA across various models reveals distinct performance patterns present in Fig. 9 and Tables 6, 7, and 8.

TABLE 5. Shows clean accuracy, adversarial accuracy, and adversarial robustness (%) of the target (ViTs and CNNs) models against the PGD perturbation method with maximum strength $\varepsilon = 0.01$ on the imagenett dataset.

Target Models	Clean Accuracy	Adversarial Accuracy	Adversaria Robustness(%)
ViT-b16	0.82	0.16	0.19
ViT-b32	0.80	0.28	0.35
ViT-L16	0.82	0.18	0.21
ViT-L32	0.75	0.29	0.38
ResNet50	0.80	0.17	0.21
VGG19	0.89	0.31	0.34
MobileNet	0.85	0.13	0.15
DenseNet	0.90	0.12	0.13

TABLE 6. Shows clean accuracy, adversarial accuracy, and adversarial robustness (%) of the target (ViTs and CNNs) models against the DFA perturbation method with maximum iteration = 3 on the imagenett dataset.

Target Models	Clean Accuracy	Adversarial Accuracy	Adversarial Robustness(%)
ViT-b16	0.82	0.13	0.15
ViT-b32	0.80	0.10	0.12
ViT-L16	0.82	0.18	0.21
ViT-L32	0.75	0.14	0.18
ResNet50	0.80	0.16	0.20
VGG19	0.89	0.09	0.10
MobileNet	0.85	0.19	0.22
DenseNet	0.90	0.10	0.11

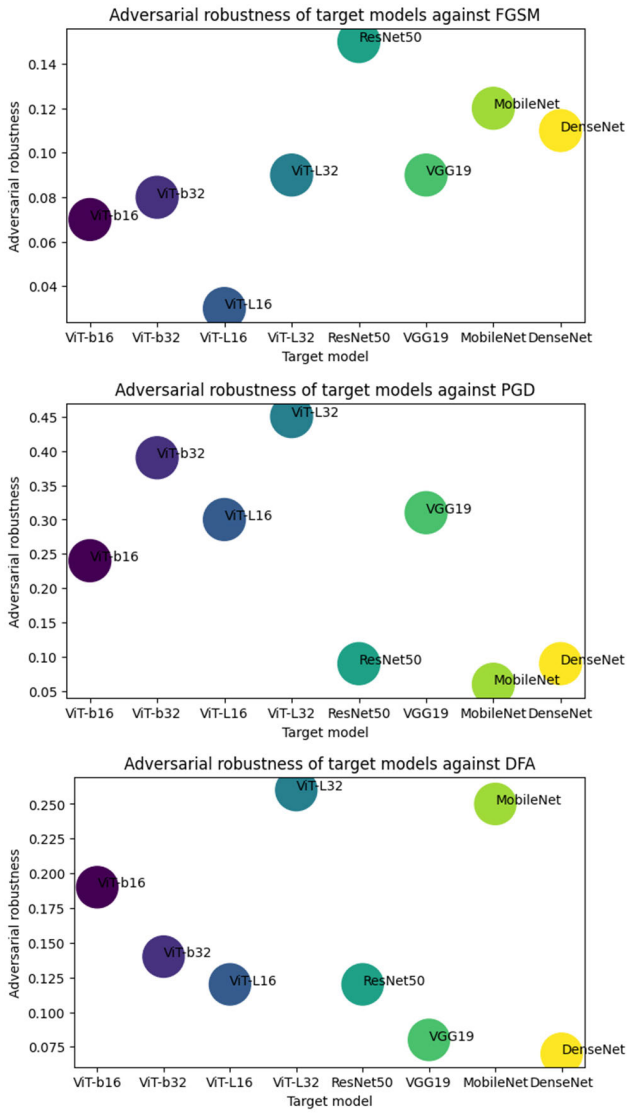


FIGURE 8. (a) The top plot compares the adversarial accuracy of target models against the FGSM perturbation method, (b) the middle plot represents the adversarial accuracy of target models against the PGD, and (c) the bottom plot shows the adversarial accuracy of target models against the DFA perturbation method on imagenett dataset.

ViT-L16 stands out with the highest robustness against FGSM, nearly reaching 0.8, while ViT-b32 excels against PGD with a robustness of 0.35. ViT models, especially the larger configurations like ViT-L16 and ViT-L32, consistently demonstrate strong robustness across all attack types. ResNet50 maintains moderate robustness across FGSM, PGD, and DFA, showcasing its balanced resilience. VGG19 shows notable robustness against PGD but is weaker against FGSM and DFA. MobileNet is highly robust against DFA, scoring around 0.22, but exhibits moderate to low robustness against FGSM and PGD. DenseNet, however, consistently displays the lowest robustness among all models tested. This analysis indicates that Vision Transformers, particularly more significant configurations, tend to have higher adversarial robustness. In contrast, traditional CNN architectures like ResNet50 and VGG19 are moderately robust, and lightweight models like MobileNet and DenseNet are less robust overall.

TABLE 7. Shows clean accuracy, adversarial accuracy, and adversarial robustness (%) of the target (ViTs and CNNs) models against the FGSM perturbation method with maximum strength $\varepsilon = 0.03$ on CIFAR10 dataset.

Target Models	Clean Accuracy	Adversarial Accuracy	Adversarial Robustness(%)
ViT-b16	0.77	0.06	0.07
ViT-b32	0.75	0.06	0.08
ViT-L16	0.82	0.03	0.03
ViT-L32	0.75	0.07	0.09
ResNet50	0.85	0.13	0.15
VGG19	0.89	0.08	0.09
MobileNet	0.77	0.10	0.12
DenseNet	0.94	0.11	0.11

TABLE 8. Shows clean accuracy, adversarial accuracy, and adversarial robustness (%) of the target (ViTs and CNNs) models against the PGD perturbation method with maximum strength $\varepsilon = 0.01$ on the CIFAR10 dataset.

Target Models	Clean Accuracy	Adversarial Accuracy	Adversaria Robustness(%)
ViT-b16	0.77	0.19	0.24
ViT-b32	0.75	0.39	0.39
ViT-L16	0.82	0.25	0.30
ViT-L32	0.75	0.34	0.45
ResNet50	0.85	0.08	0.09
VGG19	0.89	0.28	0.31
MobileNet	0.77	0.05	0.06
DenseNet	0.94	0.09	0.09

VII. DISCUSSION

The experimental results will be described in detail in the discussion section, focusing on the adversarial robustness of

TABLE 9. Shows clean accuracy, adversarial accuracy, and adversarial robustness (%) of the target (ViTs and CNNs) models against the DFA perturbation method with maximum iteration = 3 on the CIFAR10 dataset.

Target Models	Clean Accuracy	Adversarial Accuracy	Adversarial Robustness(%)
ViT-b16	0.77	0.15	0.19
ViT-b32	0.75	0.11	0.14
ViT-L16	0.82	0.10	0.12
ViT-L32	0.75	0.20	0.26
ResNet50	0.85	0.11	0.12
VGG19	0.89	0.08	0.08
MobileNet	0.77	0.20	0.25
DenseNet	0.94	0.07	0.07

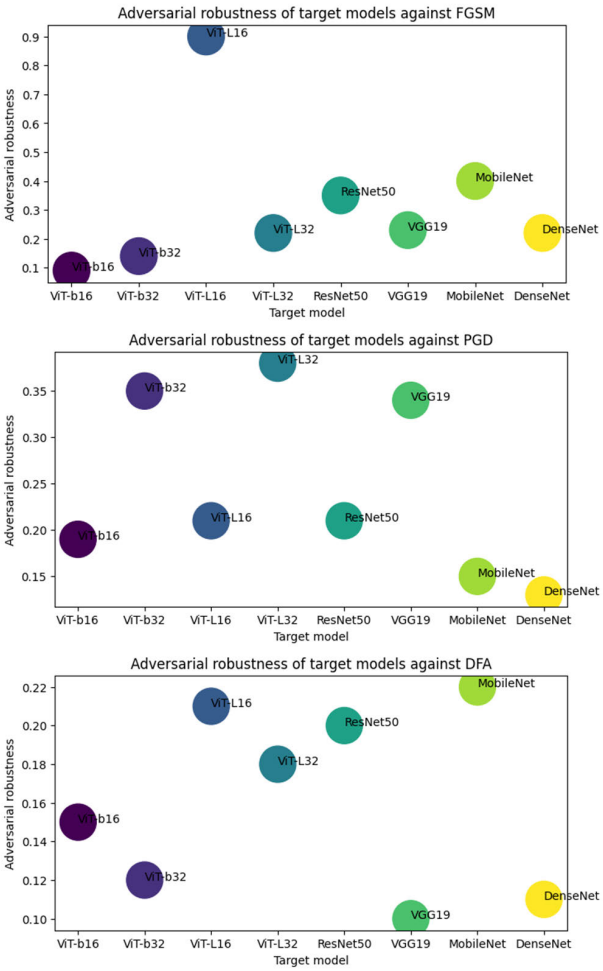


FIGURE 9. (a) The top plot compares the adversarial accuracy of target models against the FGSM perturbation method, (b) the middle plot represents the adversarial accuracy of target models against the PGD, and (c) the bottom plot shows the adversarial accuracy of target models against the DFA perturbation method on CIFAR10 dataset.

Vision Transformers (ViTs) over CNNs. It will discuss the implications of such findings, the study's limitations, and recommend some use of future research.

The analysis of the obtained experimental results shows that ViTs and CNNs have a specific antagonistic relationship in terms of adversarial robustness. Based on the results obtained, the research presents evidence that CNNs, including DenseNet and MobileNet, are more defensive against FGSM attacks. On the other hand, ViTs, especially those networks of higher capacity, such as ViT-L32 and ViT-b32, exhibit higher resistance to PGD-DFA attacks.

Specifically, DenseNet and MobileNet obtained higher adversarial robustness than ViTs while being attacked by FGSM. This implies that CNN architectures with standard layers, such as convolutional layers and feature hierarchy, might be less sensitive to straightforward gradient-based attacks such as FGSM. Out of all the investigated networks, DenseNet had the highest adversarial robustness for FGSM with a score of around 0.40, while in ViT models, the scores were much lower, starting from 0. 05 to 0.14.

All ViT models performed well against PGD attacks, and the climaxes of the robustness of ViT-L32 and ViT-b32 were 0.45. From this, we could deduce that ViTs may be more robust to iterative, gradient-based attacks because of the self-attention layer and the invariance to the global relationships in the images the model is trained on. On the other hand, CNNs such as MobileNet and DenseNet could not stave off PGD, with their scores standing at less than 0.15. It means that CNN architectures could be particularly susceptible to more complex, iterative adversarial techniques.

DFA attack results were inconclusive, while ViT models, in general, demonstrated good resilience, as apparent from the depicted robustness scores near 0 for ViT-b16 and ViT-L32 0.20. MobileNet also scored very well for robustness DFA with a score of 0.25, which represents its capability to accomplish the task even in cases when the inputs are subjected to various types of adversarial attacks.

These outcomes support the fact that adversarial robustness is a challenging problem and, at the same time, indicate that architectural choices of deep learning models have a significant impact on the defences against various adversarial perturbations. These self-attention mechanisms might explain why ViTs are less sensitive to PGD and DFA attacks than traditional models, mainly because the self-attention mechanisms are immune to several perturbations. On the same part, comparisons have shown that CNNs have high resistance to FGSM, which is a symbol of its favourable position of enduring a more straightforward attack with only a single step.

The findings from this study have several important implications for the deployment of deep learning models in security-sensitive applications:

- **Model Selection:** Depending on the adversarial threat type that is expected to be encountered, different models may be desirable. Hence, in environments that are expected to feature FGSM-like attacks, CNNs such as DenseNet might be beneficial. On the other hand, it seems that ViTs could be helpful in the case of more complex attacks like PGD or DFA.
- **Hybrid Approaches:** Because the benefits of ViTs are pretty diverse, as well as the limitations of CNNs, their combined model can increase general resilience. Such models could combine the convolutional layers to perform the extraction of features locally together with the transformers' global attention.
- **Adversarial Training:** Thus, the study confirms the necessity of using adversarial training as a defence strategy. The results also indicated that training models with a mixture of adversarial examples can make the trained models more robust regardless of the sort of attack that is adopted.

VIII. LIMITATIONS

The experiments were performed on particular databases (CIFAR10 and ImageNet to some extent). The above conclusions might be different when tested on a different dataset, especially when it is complex or possesses different characteristics. Thus, besides FGSM, PGD, and DFA, there are other types of attacks, such as C&W attacks or spatially transformed adversarial attacks, which were not examined in this work. As in most research, further research is always necessary with an enormous array of attack possibilities. The study was restricted by computational power; the experiments were performed, especially when training large ViT models. It is suggested that even more significant experiments using larger models and data can be helpful in future investigations.

IX. FUTURE RESEARCH DIRECTIONS

Therefore, more creative defence strategies that include more recent approaches of learning models or dynamic defences that work depending on the type of attack could also be invented and tested to increase model robustness further. Such concepts would enable the practical application of adversarial robustness in a driving-car context or medical imaging, for example, by testing how adversarial robustness 'translates' to realistic problems and orientating the advancement of safe AI computing. Studying new proposals that use the benefits of ViTs and CNNs may potentially result in models that are more immune to a wide range of adversarial attacks.

X. CONCLUSION

This study presents a comprehensive comparison of the adversarial robustness of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) in the context of image classification. Our experimental results indicate that while CNNs, particularly DenseNet and MobileNet, show greater robustness against FGSM attacks, ViTs, significantly larger configurations like ViT-L32 and ViT-b32, demonstrate superior resilience against iterative and sophisticated adversarial methods such as PGD and DFA. These findings suggest that the choice of model architecture significantly impacts adversarial robustness, with ViTs being more suitable for scenarios where complex adversarial threats are anticipated. The insights from this research highlight the importance of model selection based on expected adversarial conditions and pave the way for developing more robust deep learning models through hybrid architectures and enhanced adversarial training techniques. Future work should focus on exploring a broader range of adversarial attacks, testing on diverse datasets, and investigating hybrid models that leverage the strengths of both ViTs and CNNs to achieve enhanced robustness against a broad spectrum of adversarial threat methods.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number: IFP22UQU4250002DSR216.

REFERENCES

[1] V. Ashish, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[4] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[5] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, "ERNIE: Enhanced representation through knowledge integration," 2019, *arXiv:1904.09223*.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[8] A. Toppo and M. Kumar, "A review of generative pretraining from pixels," in *Proc. 3rd Int. Conf. Adv. Comput., Commun. Control Netw. (ICAC3N)*, Dec. 2021, pp. 1691–1703.

[9] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[10] N. Drenkow, N. Sani, I. Shpitser, and M. Unberath, "A systematic review of robustness in deep learning for computer vision: Mind the gap?" 2021, *arXiv:2112.00639*.

[11] A. McCarthy, "Methods for improving robustness against adversarial machine learning attacks," Ph.D. thesis, School Comput. Creative Technol., Univ. West England, Bristol, U.K., 2023.

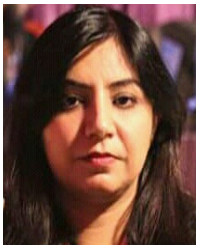
[12] A. McCarthy, E. Ghadafi, P. Andriotis, and P. Legg, "Functionality-preserving adversarial machine learning for robust classification in cybersecurity and intrusion detection domains: A survey," *J. Cybersecurity Privacy*, vol. 2, no. 1, pp. 154–190, Mar. 2022.

[13] Z. Wei, J. Chen, Z. Wu, and Y.-G. Jiang, "Adaptive cross-modal transferable adversarial attacks from images to videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3772–3783, May 2024.
[14] Y. Zhang, J. Hou, and Y. Yuan, "A comprehensive study of the robustness for LiDAR-based 3D object detectors against adversarial attacks," *Int. J. Comput. Vis.*, vol. 132, no. 5, pp. 1592–1624, May 2024.
[15] H. Liu, Z. Ge, Z. Zhou, F. Shang, Y. Liu, and L. Jiao, "Gradient correction for white-box adversarial attacks," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 11, 2023, doi: 10.1109/TNNLS.2023.3315414.
[16] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2484–2493.
[17] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," 2021, *arXiv:2102.01356*.
[18] H. Qiu, Y. Zeng, Q. Zheng, S. Guo, T. Zhang, and H. Li, "An efficient preprocessing-based approach to mitigate advanced adversarial attacks," *IEEE Trans. Comput.*, vol. 73, no. 3, pp. 645–655, Mar. 2024.
[19] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2018, pp. 634–646.
[20] D. J. Miller, Z. Xiang, and G. Kesidis, "Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks," *Proc. IEEE*, vol. 108, no. 3, pp. 402–433, Mar. 2020.
[21] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, vol. 97, 2019, pp. 7472–7482.
[22] D. Stutz, M. Hein, and B. Schiele, "Disentangling adversarial robustness and generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6969–6980.
[23] A. Ilie, M. Popescu, and A. Stefanescu, "EvoBA: An evolution strategy as a strong baseline for black-box adversarial attacks," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2021, pp. 188–200.
[24] Y.-L. Hsieh, M. Cheng, D.-C. Juan, W. Wei, W.-L. Hsu, and C.-J. Hsieh, "On the robustness of self-attentive models," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1520–1529.
[25] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT really robust? A strong baseline for natural language attack on text classification and entailment," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 8018–8025.
[26] Z. Shi and M. Huang, "Robustness to modification with shared words in paraphrase identification," 2019, *arXiv:1909.02560*.
[27] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, "BERT-ATTACK: Adversarial attack against BERT using BERT," 2020, *arXiv:2004.09984*.
[28] S. Garg and G. Ramakrishnan, "BAE: BERT-based adversarial examples for text classification," 2020, *arXiv:2004.01970*.
[29] F. Yin, Q. Long, T. Meng, and K.-W. Chang, "On the robustness of language encoders against grammatical errors," 2020, *arXiv:2005.05683*.
[30] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.
[31] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.
[32] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," 2016, *arXiv:1605.07725*.
[33] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.
[34] H. Salman, G. Yang, J. Li, P. Zhang, H. Zhang, I. Razenshteyn, and S. Bubeck, "Provably robust deep learning via adversarially trained smoothed classifiers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
[35] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333.
[36] E. Wong, L. Rice, and J. Zico Kolter, "Fast is better than free: Revisiting adversarial training," 2020, *arXiv:2001.03994*.
[37] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*.
[38] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "RobustBench: A standardized adversarial robustness benchmark," 2020, *arXiv:2010.09670*.
[39] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," in *Proc. 8th Int. Conf. Complex Netw. Appl. Complex Netw.*, vol. 1. Cham, Switzerland: Springer, 2020, pp. 928–940.
[40] K. Mahmood, R. Mahmood, and M. van Dijk, "On the robustness of vision transformers to adversarial examples," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7818–7827.
[41] Y. Qin, C. Zhang, T. Chen, B. Lakshminarayanan, A. Beutel, and X. Wang, "Understanding and improving robustness of vision transformers through patch-based negative augmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 16276–16289.
[42] H. Salman, S. Jain, E. Wong, and A. Madry, "Certified patch robustness via smoothed vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15116–15126.
[43] C. Herrmann, K. Sargent, L. Jiang, R. Zabih, H. Chang, C. Liu, D. Krishnan, and D. Sun, "Pyramid adversarial training improves ViT performance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13409–13419.
[44] M. Naseer, K. Ranasinghe, S. Khan, F. S. Khan, and F. Porikli, "On improving adversarial transferability of vision transformers," 2021, *arXiv:2106.04169*.
[45] A. Aldahdooh, W. Hamidouche, and O. Deforges, "Reveal of vision transformers robustness against adversarial attacks," 2021, *arXiv:2106.03734*.
[46] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23296–23308.
[47] S. Paul and P.-Y. Chen, "Vision transformers are robust learners," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 2071–2081.
[48] S. Tang, R. Gong, Y. Wang, A. Liu, J. Wang, X. Chen, F. Yu, X. Liu, D. Song, A. Yuille, P. H. S. Torr, and D. Tao, "RobustART: Benchmarking robustness on architecture design and training techniques," 2021, *arXiv:2109.05211*.
[49] X. Mao, G. Qi, Y. Chen, X. Li, R. Duan, S. Ye, Y. He, and H. Xue, "Towards robust vision transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12032–12041.
[50] K. Jeeveswaran, S. Kathiresan, A. Varma, O. Magdy, B. Zonooz, and E. Arani, "A comprehensive study of vision transformers on dense prediction tasks," 2022, *arXiv:2201.08683*.
[51] S. Shleifer and E. Prokop, "Using small proxy datasets to accelerate hyperparameter search," 2019, *arXiv:1906.04887*.
[52] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2017, *arXiv:1712.09665*.
[53] A. Homaifar, C. X. Qi, and S. H. Lai, "Constrained optimization via genetic algorithms," *Simulation*, vol. 62, no. 4, pp. 242–253, Apr. 1994.
[54] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
[55] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.
[56] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.

KAZIM ALI received the Ph.D. degree from the University of Central Punjab, Lahore, Pakistan. He has more than 14 years of experience as a Computer Science Lecturer with the Government Education Department, Punjab, Pakistan. He has one year of experience as a Software Instructor with Punjab Vocational Institution Sukheke, Mandi Hafizabad. He has research and development experience in machine learning, deep learning, computer vision, and natural language experience fields.

MUHAMMAD SHAHID BHATTI received the master's degree in computer science from the University of Central Punjab, Lahore, Pakistan. He is currently an esteemed Assistant Professor with COMSATS University Islamabad. He is distinguished for his profound expertise in computer vision, a testament to his expansive knowledge of machine learning. His academic and research endeavors encompass many subjects, such as computational intelligence and data science, data visualization, and digital image processing. Throughout his illustrious career, he has made significant scholarly contributions, having penned several notable research articles, that reflect his dedication and passion for the ever-evolving realm of technology. Beyond his writings, he remains actively involved in cutting-edge research across the aforementioned fields, continuously striving for innovation, and enriching academic discourse.

ATIF SAEED is currently an Active Computer Scientist. He is also an Active Member of the Computer Network and Research Group, COMSATS University Islamabad, Lahore Campus, and Lancaster University U.K. He is recognized as an "HEC Approved Supervisor." He has published over 15 peer-reviewed articles in the field of networks and distributed systems. His research interests include the complexity of cyber security and hacking (cloud computing, data centers, the Internet of Things, and networks). This entails the analysis, modeling, and exploitation of emergent network and cloud system phenomena to propose novel techniques to enhance system security, dependability, resource management, designing of new secure network architecture, and energy efficiency.



ATIFA ATHAR is currently an Assistant Professor of computer science. Her research interests include artificial intelligence, cognitive machines, artificial neural networks, computational intelligence, and fuzzy logic.

MOHAMMED A. AL GHAMDI received the bachelor's degree (Hons.) in computer science from King Abdul Aziz University, Jeddah, Saudi Arabia, in 2004, the master's degree (Hons.) in internet software systems from Birmingham University, Birmingham, U.K., in 2007, and the Ph.D. degree in computer science from the University of Warwick, U.K., in 2012. Since 2012, he has been with the Department of Computer Science, Umm Al-Qura University, Makkah, Saudi Arabia, as an Assistant Professor and then as an Associate Professor. He is currently the Founder and the Scientific Chair of data and artificial intelligence with Umm Al-Qura University. He has authored over 50 papers in international conferences and journals, such as IEEE Systems Journal, IEEE Transactions on Engineering Management, IEEE Access, *Computers, Materials and Continua* (CMC), IEEE International Conference on Scalable Computing and Communications, and International Conference on Cloud Computing and Services Science. His research interests include machine learning, data analysis, AI, cloud computing, and cybersecurity.

SULTAN H. ALMOTIRI received the B.Sc. degree (Hons.) in computer science from King Abdulaziz University, Saudi Arabia, in 2003, and the M.Sc. degree in internet, computer, and system security and the Ph.D. degree in wireless security from Bradford University, U.K., in 2006 and 2013, respectively. He was the Chairman of the Computer Science Department, Umm Al-Qura University, Saudi Arabia; the Vice Dean of eLearning and distance education; and the Chief Cybersecurity Officer in general administration of cybersecurity with Umm Al-Qura University. He is currently a member of the Scientific Council, Umm Al-Qura University, and an Associate Professor with the Cyber Security Department, Faculty of Computing. His research interests include cyber security, cryptography, AI, machine learning, eHealth, eLearning, the IoT, RFID and wireless sensors, and image processing.

SAMINA AKRAM is currently a Lecturer of computer science with the FoIT, University of Central Punjab, Lahore, Pakistan. Her research interests include machine learning, artificial neural networks, artificial intelligence, computational intelligence, fuzzy logic, and medical image processing.